QSAR Modeling of Human Serum Protein Binding with Several Modeling Techniques Utilizing Structure−Information Representation

Joseph R. Votano,*[,⊥] Marc Parham,[⊥] L. Mark Hall,[†] Lowell H. Hall,[‡] Lemont B. Kier,[§] Scott Oloff,[||] and Alexander Tropsha[||]

ChemSilico LLC, 48 Baldwin Street, Tewksbury, Massachusetts 01876, Hall Associates Consulting, 2 Davis Street, Quincy, Massachusetts 02170, Department of Chemistry, Eastern Nazarene College, Quincy, Massachusetts 02170, Department of Medicinal Chemistry, School of Pharmacy and The Centre for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, Virginia 23298, and Laboratory of Molecular Modeling, Division of Medicinal Chemistry and Natural Products, School of Pharmacy, University of North Carolina, Chapel Hill, North Carolina 27599

Received December 14, 2005

Four modeling techniques, using topological descriptors to represent molecular structure, were employed to produce models of human serum protein binding (% bound) on a data set of 1008 experimental values, carefully screened from publicly available sources. To our knowledge, this data is the largest set on human serum protein binding reported for QSAR modeling. The data was partitioned into a training set of 808 compounds and an external validation test set of 200 compounds. Partitioning was accomplished by clustering the compounds in a structure descriptor space so that random sampling of 20% of the whole data set produced an external test set that is a good representative of the training set with respect to both structure and protein binding values. The four modeling techniques include multiple linear regression (MLR), artificial neural networks (ANN), k-nearest neighbors (kNN), and support vector machines (SVM). With the exception of the MLR model, the ANN, kNN, and SVM QSARs were ensemble models. Training set correlation coefficients and mean absolute error ranged from $r^2 = 0.90$ and MAE = 7.6 for ANN to $r^2 = 0.61$ and MAE = 16.2 for MLR. Prediction results from the validation set yielded correlation coefficients and mean absolute errors which ranged from $r^2 = 0.70$ and MAE = 14.1 for ANN to a low of $r^2 = 0.59$ and MAE = 18.3 for the SVM model. Structure descriptors that contribute significantly to the models are discussed and compared with those found in other published models. For the ANN model, structure descriptor trends with respect to their affects on predicted protein binding can assist the chemist in structure modification during the drug design process.

Introduction

Most drugs bind reversibly with varying degrees of association to human plasma proteins: serum albumin (HSA), alpha-1-acid glycoprotein (AGP), and lipoproteins. The degree of binding, expressed as the percent bound (%PB), varies from 0% to 100%. Reported association constants[1,2] range from $\sim 10^{-3}$ to $\sim 10^{-10}$ M^{-1}. For AGP a high value of 5×10^{-6} M^{-1} has been reported for HIV protease inhibitors.[3]

Since the drug−protein complex in the plasma acts as a reservoir for the drug, the %PB is an important parameter in pharmacokinetic profiling. For this reason protein binding influences many aspects of ADME/Tox properties such as metabolism, excretion, and in vivo activity. The latter is especially true when a drug candidate possesses both an undesirable physicochemical property (e.g., poor aqueous solubility) as well as an undesirable pharmacological property (e.g., a high effective concentration requirement). In this case, the %PB must have a low to moderate value for the potential candidate to have a successful therapeutic consequence in clinical trials. In the reverse situation, where %PB is high (>99.9%) and affinity for HSA is larger than that of the receptor targeted by the drug, the volume of distribution of the drug becomes highly restrictive; making it likely the candidate would be unsuitable as therapeutic agent. For reasons such as these, much research attention is directed toward managing the protein binding of drug candidates.

In human serum plasma proteins, the primary constituent is HSA, with a lesser amount of AGP and an even smaller amount of lipoproteins.[4] The plasma concentration of HSA is around 600 μM for this 66 kDa globular protein, which consists of three, very similar 3-D structural domains, designated as I, II, and III. Each domain possesses two subdomains, A and B. High-resolution X-ray structures have identified eight subdomains: eight fatty acid binding sites within all six A and B subdomains.[5,6] By contrast, drug-like compounds have been suggested to bind at either of two high affinity sites. Site-I in the IIA subdomain is commonly referred to as the warfarin site.[5,7,8] Site-II, in subdomain IIIA, is called the diazepam site.[5] Sites I and II are quite similar in size and shape, possessing elongated hydrophobic pockets with polar residues at the mouth and side walls, close to the cavity entrance. The pockets are lined with hydrophobic residues: Phe, Trp, Ile, Ala, and Leu. At the entrance of Site I, Arg, His, and Lys can undergo electrostatic and/or H-bonding interactions with negatively charged groups or H-acceptors on drug-like compounds. Site-I binds acidic and neutral compounds. Arg and Tyr at the cavity entrance of Site II bind entities that are neutral or basic at pH 7.4.[9] In light of that fact that the specific binding mode of binding to HSA has been verified for a limited number of drugs, it is certain that additional binding sites for molecule compounds to HSA do exist. Paclitaxel, a potent antitumor agent, has been shown to have more than six HSA binding sites, with the initial site having high affinity binding (i.e., $K_d = 2 \times 10^{-6}$ M^{-1}).[10] Two general

* Corresponding author. Tel 1-978-501-0633, Fax: 1-781-275-5197, e-mail: jvotano@chemsilico.com.
[⊥] ChemSilico LLC.
[‡] Eastern Nazarene College.
[†] Hall Associates Consulting.
[§] Virginia Commonwealth University.
[||] University of North Carolina.

10.1021/jm051245v CCC: $33.50
Published on Web 11/07/2006

anesthetics, propofol and halothane, are known to bind to several fatty acid sites.[11] A quantitative assessment of the binding modes (subdomains involved, number of sites, types of interactions) for the great majority of HSA bound compounds is lacking. Nonetheless, the need to develop in silico models to predict %PB for early-stage drug candidate selection remains important.

In recent years, published QSAR models for HSA binding were based primarily on small datasets, usually less than 350 compounds. In this study we examined over 1000 drug and drug-like compounds with reported %PB values encountered with plasma proteins. To our knowledge, the model reported here is based on the largest human serum protein binding data set taken from the literature. In this work we attempt not only to provide prediction for new chemical entities (NCEs) but also to elucidate the important physicochemical properties, structural attributes, and substituent groups that contribute to %PB. Four different types of quantitative-structure−activity relationship (QSAR) models were developed in this present investigation to assess the commonality of descriptors that might exist among these models and to assess their robustness for NCE prediction (validation set). In addition, to compare descriptors found here with reported results from several other studies.

Materials and Methods

Sources of Compounds and Their Attributes. Data on the percent fraction of compounds bound to plasma proteins (%PB) came from a variety of sources.[12−17] A clustering technique was used to sort compounds into small groups with similar structures. If the reported experimental value for a compound was substantially different from the values of other compounds in the same cluster, values were checked for the presence of a data error. Common data errors included: reported fraction unbound instead of fraction bound, value not measured in human plasma, or the value reported was for the primary metabolite. Of the compounds selected, 418 (41%) had two or more reported values. If two or more reported values for the same compound differed by 30% or more, the compound was excluded; otherwise they were averaged. Several other classes of compounds excluded were proteins, organometallics, and those reported to show a dose or time dependency. The number of unacceptable compounds found in various data sources was 103 including 68 duplicates. Twelve compounds selected for the 1008 dataset had no quantitative value for %PB. These were reported as negligible, poor, or high %PB and were assigned %PB values of 2, 25, and 80%, respectively. Twenty-two compounds with fuzzy assignments were found among the high binders. Those reported with %PB > 90% and > 99.5 were assigned a %PB of 95% and 100%, respectively.

All chemical structures used in molecular descriptor computations were in the neutral form except 25 compounds with a permanent positive charge (e.g., quaternary amines). Table 1 gives important compound attributes for the 1008 drug dataset. Approximately 97% of drug compounds had one or more ring structures with an average of approximately three rings per structure; 54% had fused rings, and 26% had one or more heteroaromatic rings. On average each compound had nine rotatable bonds. As shown in Table 1, approximately one-quarter of the drugs have a carboxylic acid group, 26% contain a halogen atom, 61% contained at least one amine, 35% have an amide group, and 39% have a hydroxyl group. As expected, the molecular weight, number of hydrogen bond donors and acceptors, and total polar surface area of these compounds are consistent with drug-like compounds. An approximate gauge of the chemical diversity of 1008 compounds (results not given) is revealed by a principal component analysis based on 115 structure descriptors (molecular connectivity and atom-type E-state indices). The first two PCA components explain only 28% of the variance, and the first nine components explain 59% of variance, indicating a high level of chemical diversity among these compounds.

Table 1. Compound Attributes in Train and Validation Sets

structure attribute	average[a]	number[b]	percent
Ring[c]	2.84	978	97.0%
N-heteroaromatic ring[d]	0.26	263	26.1%
nonheteroaromatic ring[d]	0.54	549	54.5%
nonaromatic ring[d]	0.16	166	16.5%
fused ring system	0.54	543	53.9%
rotatable bonds	9.2	1008	100%
$-CO_2H$	0.26	261	25.9%
$-NO_2$	0.03	33	3.3%
amines[e]	1.7	431	42.8
amides[f]	0.67	391	38.8%
−OH	0.75	358	35.5%
halogens[g]	0.41	258	25.6%
average NumHBa	6.63	1007	99.9%
average NumHBd	2.12	847	84.0%
average MW[h]	362.4	1008	100%
TPSA[i]	96.4	1008	100%

[a] Average value for attribute in dataset. [b] Number of compounds with specified attribute. [c] Compound contains a ring structure. [d] Type of ring structure. [e] Primary, secondary, and tertiary amines. [f] Amides and sulfonyl amides. [g] All Halogens F, Cl, Br, and I. [h] MW = Molecular weight. [i] TPSA = total static polar surface area of O, N, P, and S along with associated hydrogen atoms.

Molecular Descriptor Selection. An initial set of 628 topological structure descriptors were computed by ChemSilico software[18] and reduced to a set of 180, using the criterion that at least 3% of the descriptor values must have nonzero variance (nonzero in most cases for the 1008 compounds). The descriptors include molecular connectivity chi indices, E-State indices of the atom-, bond-, and group-type as well as atom- and group-type hydrogen E-State descriptors, kappa shape indices, and several binary indicators (e.g., presence of aromatic ring, types of amides, acids). Counts of atoms, groups, or fragments were not included for modeling. Predicted LogP was calculated using CSLogP[18] and included as a bulk property descriptor, resulting in 181 total descriptors available in the initial set. This initial set of 181 descriptors was further reduced by the various selection routines implemented in conjunction with the four modeling algorithms investigated in this study. The approach employed in this investigation for encoding molecular structure in topological descriptors is referred to as the structure−information representation whose significance for modeling biologically important properties has been discussed.[19]

Selection Process for Train and Validation Compound Sets. To select a training set and a validation set of compounds, Ward's hierarchical clustering was performed using MDL QSAR software.[20] The set of descriptors used for clustering consisted of 115 topological structure descriptors containing only molecular connectivity chi indices and atom-type E-State descriptors. Ten clusters were produced. The average cluster size was 112 compounds, excluding the two smallest clusters, each containing two compounds. The validation set (NCEs) was created by random selection of 20% of the compounds from each cluster to yield 200 compounds. The four compounds in the two smallest clusters were assigned to the training set. The compounds in the external validation set were used to determine the predictive capabilities of the four QSAR models developed in this study but did not contribute to any phase of model development.

Model Development

QSAR models were developed using four different modeling algorithms: multiple linear regression (MLR), artificial neural networks (ANN), k-nearest neighbors (kNN), and support vector machine (SVM). Each modeling procedure started with the same initial set of 181 descriptors (180 topological structure descriptors together with predicted logP) as independent variables, 808 compounds for train/test set, and 200 for external validation.

MLR Model Development. MLR analysis was accomplished with JMP v5 (SAS Institute Inc., Cary, NC) on the 808 compound train set. In the MLR modeling process, a step-

forward selection process was conducted until the r^2 reached a semi-plateau where changes in r^2 were less than 0.2% upon adding another variable, under the constraint that the number of descriptors was always less than $\sqrt{N} + 5$ (N = number of compounds). Removal of potentially redundant descriptors was accomplished using the criterion that the inter-correlation r^2 for all pairwise variables be less 0.90. The final model contained 30 descriptors for the best r^2 value. Ranking of the final set of 30 descriptors, to indicate relative importance, was determined by a leave-one-out approach and ranked by the sum of the residues squared (RSS) in the absence of the descriptor. A 100-fold randomization of %PB values was performed with r^2 computed for each case, yielding an average r^2 less than 0.03 for the MLR model. The results of this randomization process indicate that the model is different from an equation based on random numbers, suggesting that significant information is contained in the model.

ANN Model Development. For the Artificial Neural Network (ANN) analysis, the 808 compound train/test set, designated the principal set, was randomly split into 90% for training and 10% as a (internal) test set. This process was repeated 10 times to produce 10 mutually exclusive train/test sets (or folds) of the data. The selection was carried out such that each compound in the principal set appeared in a test set only once and was used for training nine times. A standard back-propagation neural network was used for this study. The network contained no more than nine hidden neurons and utilized the backward elimination approach[21,22] for descriptor selection which has been adapted from traditional linear regression methods.

Each training set was processed separately with the neural network algorithm, using the test set to prevent over fitting. Each model was applied to the corresponding (internal) test set to calculate q^2, which is the r^2 value for all instances in which the data was withheld from the modeling process. This multiple selection process leads to a set of 10 models with predicted values which were averaged, called an ensemble model. The average value of 10 neural nets, the ensemble model, gives the predicted %PB value of a compound. Ranking of descriptors with respect to their importance in the model was determined as the ratio of the difference in RSS (sum of squares of residuals) in the presence and absence of the variable, divided by the smallest difference (the least important variable in the train-test set), using an average RSS values from all ten ANN models. Using this approach, the variables judged to be noncontributory are pruned during the 10-fold cross-validation in a backward stepwise manner until the r^2 declined two consecutive times by more than 0.02 r^2 units. The last model, just prior to this drop in r^2, was selected as one with the optimal descriptor subset.

As a general rule, we considered only models with an absolute value of $q^2 = 0.50$ as the minimum cut off value for q^2 (with corresponding higher values for r^2). Typically, we observed that r^2 values were greater than the corresponding q^2 by 0.1 to 0.3 units. By this backward elimination process the initial starting set of 181 descriptors was reduced to 33 in this ANN analysis. The relative importance of each eliminated variable is based on its contribution across the entire train/cross-validation sets by calculation of r^2.

Since the training r^2 is used to select the variables, it does not provide a reliable assessment of the predictive accuracy of the overall algorithm. This task is reserved for the independent external validation set of 200 compounds, which was not used to generate an algorithm or select an optimum subset of inputs during the backward elimination process.

Model Developed by kNN-QSAR. The kNN QSAR method[23] employs the kNN pattern recognition principle[24] and a variable selection procedure. Briefly, a fixed size subset of descriptors is selected randomly in the beginning of the calculations. The model is built using this random descriptor selection with leave-one-out (LOO) cross-validation, where each compound is eliminated from the training set and its biological activity is predicted as the average activity of the k most similar molecules ($k = 1-5$). The similarity is characterized by the Euclidean distance between compounds in multidimensional descriptor space. A method of simulated annealing with the Metropolis-like acceptance criterion[23] is used to sample the entire descriptor space to converge on the subset of the same size which affords the highest value of LOO r^2 (q^2). The descriptor subsets of different sizes are optimized using this procedure to arrive at a variety of models with acceptable q^2 greater than certain threshold (0.6 was selected as the default threshold). The training set models with acceptable q^2 are then validated on external test sets to select the 10 predictive models, the ensemble model, as discussed above. Further details of the kNN method implementation, including the description of the simulated annealing procedure used for stochastic sampling of the descriptor space, are given elsewhere.[23]

The original kNN method[23] was enhanced by using weighted molecular similarity. In the original method, the activity of each compound was predicted as the algebraic average activity of its k-nearest-neighbor compounds in the training set. In general, however, the Euclidean distances in the descriptor space between a compound and each of its k nearest neighbors are not the same. Thus, the neighbor with the smaller distance from a compound is given a higher weight, with exponential dependence on distance, in calculating the predicted activity as follows:

$$w_i = \frac{\exp(-d_i)}{\sum_{i=0}^{i=k} \exp(-d_i)}$$

$$y = \sum w_i y_i$$

Here d_i is the Euclidean distance between the compound and its k nearest neighbors, k is the number of nearest neighbors, w_i is the weight for every individual nearest neighbor, and y_i is the actual activity value for ith nearest neighbor. In summary, the kNN algorithm generates both an optimum k value and an optimal subset of descriptors, which afford a QSAR model with the highest value of q^2. This modified algorithm was also applied recently to the kNN QSAR modeling of anticonvulsant agents.[25]

Model Development by SVM. The support Vector Machine technique (SVM) was developed by Vapnik[26] as a general data modeling methodology where both the training set error and the model complexity are incorporated into a special loss function that is minimized during model development. The methodology allows one to regulate the importance of the training set error versus the model complexity to develop the optimum model that best predicts a test set. In later developments, SVM was extended to afford the development of SVM regression models for datasets with noninteger activities and used for QSAR development.

We have implemented the SVM method for QSAR modeling to construct a 10 member ensemble model. Each SVM member was developed as follows: let m be the number of points representing the training set compounds with known biological activity in an n-dimensional descriptor space. The problem is

to generate a hyper-surface in the descriptor-activity (n+1) dimensional space that relates descriptor values to the biological activities. Thus, the biological activity of any compound can be predicted from its descriptors by placing the point corresponding to this compound on this hyper-surface.

Given a training set of instance-label pairs (x_i, y_i), $i = 1, ..., m$ where x_i, ϵ, R^n are the descriptors that describe each compound and y_i is the biological activity (e.g., %PB) of each compound, the sought correlation between structure and activity can be represented as $y_i = f(x_i)$. For simplicity, we will define $f(x_i)$ to be a linear function of the form

$$f(x_i) = \langle \omega_i, x_i \rangle + b$$

where ω is the coefficient vector of the linear function and b is the bias. One major goal of any regression algorithm is to minimize the errors between the predicted and the actual values as defined by ξ_i in the following equation:

$$|y_i - (\langle \omega_i, x_i \rangle) + b| = \xi_i$$

As a means of regulating generalization of the algorithm, SVM utilizes the following constraint to solve the optimization problem: with the constraint:

$$\min_{\omega,b,\xi}^{?} \frac{\omega^T \omega}{2} + C \sum_{i=1}^{m} \xi_i$$

$$|y_i - (\omega^T \Phi(x_i) + b| = \xi_i$$

whereas the training vectors x_i are mapped into a higher dimensional space by a kernel function Φ. Then the SVM algorithm finds a linear correlation between the actual activity and this higher dimensional space $\Phi(x_i)$. The quantity C (> 0) is the penalty parameter of the error term that controls the weight between the two terms in the SVM optimization problem. During optimization, the relative weights are assigned to each descriptor whereby a large number of the descriptors are given a weight of zero to minimize the value of the loss function. However, a small number of descriptors are given a nonzero weight and the absolute value of that weight implies the relative significance that descriptor has on activity prediction.

In many cases the binding activities may contain small errors or the kernel function may not be capable of perfectly representing the training compounds in a simplified manner. As a means of inhibiting the algorithm from producing an overly complicated training set correlation that would not accurately predict a test set, we included a slack variable, ϵ. This slack variable is a threshold of prediction error for any compound's activity before the algorithm is penalized for a poor prediction. Beyond the boundary ϵ the algorithm is penalized by the value of $\xi_i - \epsilon$. When combining the SVM optimization problem defined with a linear kernel, the following SVM loss function is obtained:

$$_{\min}\text{loss} = \frac{||\omega||}{2} + C \frac{i=1}{m} \begin{bmatrix} 0 \text{ if } \xi_i < \epsilon \\ \xi_i - \epsilon \text{ if } \xi_i > \epsilon \end{bmatrix}$$

The nature of SVMs requires one to specify a priori the values of C and ϵ since it is not known beforehand which values may work best for the dataset; thus, a parameter search must be performed. The goal is to identify good values of C and ϵ such that the model can accurately predict unknown data (i.e., testing data). In most circumstances, the highest training accuracy does not yield the best accuracy on a test set. Therefore, the optimum C and ϵ values are commonly selected based on the values that give the best test set results.

Table 2. Results of Four QSAR Models for Training and Validation Datasets

model	number	R^2	MAE[a]	RMSE[b]	descriptors
Training Set Statistics					
ANN	808	0.90	7.6	10.8	33
kNN	808	0.62	15.6	20.9	29
MLR	808	0.61	16.2	21.0	30
SVM	808	0.62	16.2	21.7	61
Validation Set Statistics					
ANN	200	0.70	14.1	18.6	33
kNN	200	0.59	16.7	21.8	29
MLR	200	0.59	17.2	21.8	30
SVM	200	0.59	18.3	23.3	61

[a] MAE, mean absolute error, $= 1/N \times \sum (|\%PB_{exp} - \%PB_{pred}|)$ where N = number of compounds. [b] RMSE is root-mean-square error.

In many cases we use a "grid-search" on C and ϵ to identify the best parameters. There are several advanced methods which can save computational cost by estimating the best parameters. There are two reasons why we preferred a simple grid-search approach. First, unlike alternative methods which use approximations or heuristics, a grid-search allows for an exhaustive parameter search and does not have a convergence problem due to local minima. Second, the computational time to find good parameters by a grid-search is not much longer than the time required by advanced methods since there are only two optimization parameters. Furthermore, the grid-search can be easily parallelized because each parameter is independent. Many of the advanced methods for parameter estimation are iterative processes, e.g. walking along a path, which is difficult for parallelization.

For large datasets, a complete grid-search may be overly time-consuming; therefore, we commonly use first a coarse grid on a subset of available data. A user may randomly choose a subset of the dataset, conduct a grid-search using those compounds, and then do a fine-tuned grid-search on the complete dataset over the parameter value ranges that exhibited the best results.

Results

Statistical Information on the QSAR Models. Four different QSAR algorithms were employed in this investigation. MLR was used as a baseline regression method along with three machine learning algorithm approaches, ANN, kNN, and SVM, to investigate their performance in model development based on fitting a training set followed by predictions on an external validation set. Statistical results from prediction of the training set (808 compounds) and the validation set (200 compounds) by each of the four models are summarized in Table 2. For the training set, the ANN model gave the lowest mean absolute error (MAE) equal to 7.6 using 33 descriptors, followed by the kNN model with MAE = 15.6 using 29 descriptors. Surprisingly, the MLR and SVM models gave essentially the same statistical outcomes: MAE =16.2 and 16.6, respectively. The SVM ensemble model employed 61 variables. Another characteristic of models is the ratio of observations to input variables (i.e., structure descriptors and logP). For three of the models that ratio is 24 (ANN) and 27 (kNN and MLR). However, the SVM model required nearly twice as many descriptors as the other models, leading-to-a ratio of 13. Generally a ratio greater than 10 is considered reasonable for QSAR models.

The most important criteria of how robust a QSAR model is, are differences in the predicted versus experimental values coming from the use of a compound validation set. Of the four QSAR models, ANN and kNN performed best with MAE =

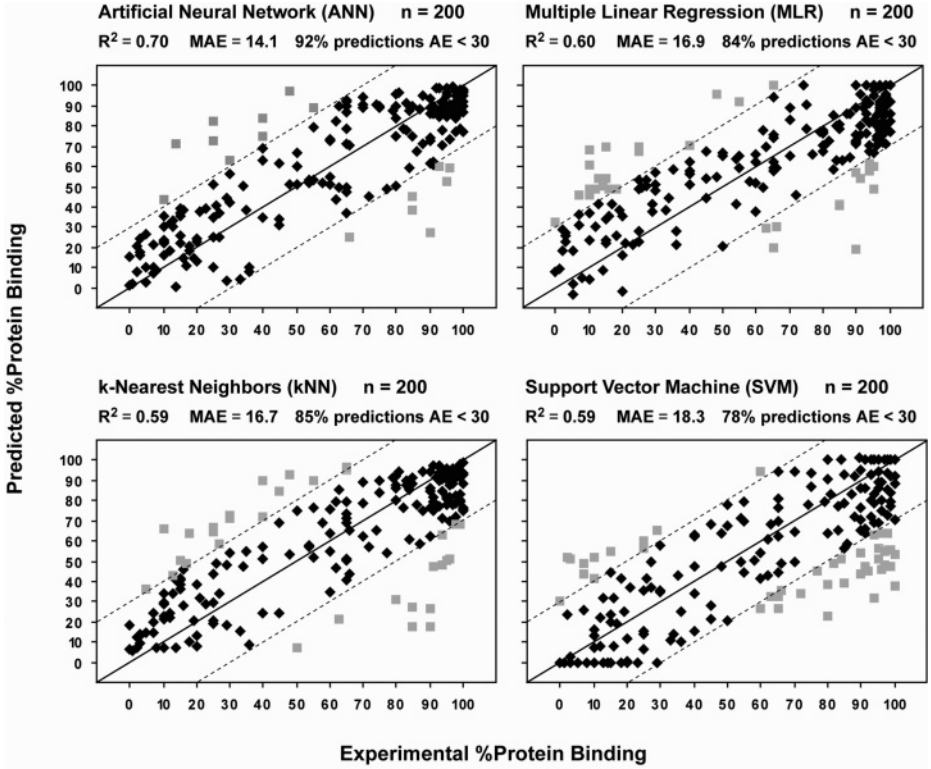


Figure 1. QSAR modeling results for a 200 compound validate set for four models; ANN = artificial neural net, kNN = k-nearest neighbor, MLR = multiple linear regression, SVM = support vector machine. Dotted diagonals are 30% off-sets from experimental PB values. Gray boxes represent predictions outside the ± 30% range.

14.1 and 16.7, respectively. The MLR and SVM models were quite similar in their predictive ability, differing by only 6% in their MAE values; MAE = 17.2 and 18.3, respectively (See Table 2). Considerable differences did exist in the predictive capabilities among the models as revealed in the distribution of predictions for the 200 compound validation set, shown in Figure 1. These differences may be highlighted by subdividing the validation set into three subsets based on the range of %PB values and then examining the prediction quality within each subset. The lower range is defined as %PB < 15% (14% of the validation set), a midrange from 15 to 85% (48% of the validation set), and an upper range of compounds with %PB >85% (38% of the validation set). The MAE values for the ANN model were 14, 17, and 10 in these three %PB ranges. The corresponding values for the kNN model are 15, 19, and 14. In the bottom range, the MLR and SVM models yielded MAE = 25 and 16, respectively. In the mid and top ranges the MAE values for MLR and SVM were 17 and 19, and 14 and 19, respectively. It is evident the models are not balanced in their MAE values across these three subsets. The ANN model does better at the high %PB range than any of the other QSAR models and somewhat better in the low to middle ranges. The same unbalanced predictions also occur when these subset ranges are expanded or contracted by 15% in their PB values.

Table 3 summarizes results from all four models for the external validation set: experimental PB_{exp}, predicted %PB values, absolute error (AE), and the calculated logP values (CSLogP) for each of the 200 compounds. All models agreed within less than 15% in their predicted %PB values for only 34% of the validation compounds. Best agreement, less 15% difference in their predicted %PB values, was found among the ANN and kNN models on 158 compounds in 200 member validation set. However, it is apparent from Table 3 the spread in AE values from the four models for a given compound can be small or large; e.g., good predictive agreement among the models for arebekacin and poor agreement for acenocoumarol. Furthermore, the range in AE values, 0% to 72%, prompted us to examine the worst predicted compounds for each model. We computed r^2 and MAE for all four models upon removal of 5% of the compounds with largest AE values, 10 for each model. On average, r^2 increased substantially (by 17%) and MAE declined significantly (by 11%) for the models when 5% of the data with the largest residuals was removed. Examination of these compounds (where AE ranged from 40% to 72% among the models) revealed that three compounds, ambroxol, niridazole, and desbrisoquine, were common to the ANN, MLR, and kNN models. These three compounds plus tilidine, pravastatin, cilazapril, and zonisamide were common to the ANN and MLR models. For the SVM model, only two compounds, nafarelin and buprenorphine, were found in common only with the kNN model. These model dependent outliers have very little structural similarity; so unless the reported %PB values are in error, the models could not predict binding values for these compounds within 40% of their experimental ones.

Important Structure Descriptors. A summary of the descriptors ranked as the top 20 in importance (19 topological indices plus logP) together with their frequency of occurrence in the train/test dataset are given in Table 4 for all four QSAR models. The mean trends in descriptor values with %PB are also given for two models, ANN and MLR. Here frequency of occurrence is defined as the percentage of compounds for which the descriptor is present, nonzero. A global descriptor is defined as one for which the frequency > 90%. Atom-type E-state descriptors, signified by a capital S, and bond-type indices, signified by e, are the predominant descriptors in all models. These two descriptor classes as a percent of the total descriptors found in a model ranged from 95% in the ANN model to 50% in the kNN model with the remaining indices being molecular connectivity descriptors. LogP, the most important descriptor in all models, had a wide range of values from −3 to 7 with

Table 3. Results for QSAR Models on 200 Validation Compounds for Percent Serum Protein Binding (%PB)

compound	CSLogP [a]	PBexp[b]	PBpred[c] (ANN)	AE[d]	PBpred (kNN)	AE	PBpred (MLR)	AE	PBpred (SVM)	AE
17-β-estradiol	3.33	98	95	3	93	5	83	15	94	4
17-hydroxyprogesterone	2.84	85	90	5	88	3	75	10	93	8
acenocoumarol	1.71	98	97	1	72	26	87	11	85	13
albendazole sulfoxide	1.76	70	89	19	62	8	66	4	65	5
alfadolone	1.86	30	63	33	73	43	57	27	58	28
amantadine	2.10	66	25	41	44	22	29	37	35	31
ambroxol	0.05	90	28	62	27	63	22	68	44	46
amdinocillin	1.10	14	71	57	41	27	48	34	0	14
amisulpride	1.16	16	15	1	46	30	40	24	32	16
amithiozone	1.24	95	53	42	50	45	53	42	56	39
amitriptyline	3.93	95	85	10	92	3	83	12	86	9
amoxicillin	−1.49	19	15	4	21	2	34	15	25	6
aprobarbital	1.16	62	44	18	47	15	48	14	43	19
arbekacin	−3.94	8	9	1	7	1	0	8	0	8
azithromycin	3.77	35	39	4	55	20	53	18	43	8
baclofen	−0.68	30	42	12	71	41	52	22	35	5
barbital	0.75	10	36	26	29	19	43	33	13	3
bleomycin	4.80	10	30	20	30	20	63	53	16	6
bopindolol	2.57	65	93	28	71	6	67	2	77	12
brotizolam	2.90	90	93	3	77	13	100	10	95	5
bunazosin	2.18	97	91	6	78	19	77	20	100	3
buprenorphine	4.20	96	59	37	51	45	87	9	46	50
calcifediol	4.16	80	96	16	92	12	80	0	100	20
capecitabine	1.47	30	56	26	54	24	51	21	35	5
carbenoxolone	3.89	100	87	13	93	7	100	0	53	47
carbidopa	−2.33	36	10	26	9	27	28	8	25	11
carbimazole	0.15	7	10	3	15	8	37	30	49	42
cefdinir	−1.87	65	47	18	41	24	17	48	32	33
cefoperazone	1.08	91	62	29	80	11	78	13	66	25
cefoxitin	0.42	72	45	27	57	15	44	28	34	38
cefprozil	−0.89	40	35	5	24	16	49	9	34	6
cefsulodin	0.11	23	39	16	29	6	17	6	0	23
ceftizoxime	−0.58	29	44	15	48	19	45	16	65	36
cetirizine	2.60	93	60	33	97	4	80	13	80	13
cevimeline	1.52	15	26	11	41	26	25	10	24	9
cilazapril	0.65	25	73	48	44	19	71	46	15	10
cilostazol	3.30	97	99	2	88	9	93	4	100	3
cladribine	0.16	20	13	7	13	7	34	14	12	8
clindamycin	2.59	94	71	23	71	23	55	39	32	62
clomipramine	5.25	97	100	3	96	1	96	1	56	41
clonidine	1.32	25	35	10	64	39	50	25	36	11
clorazepic acid	3.26	91	90	1	97	6	98	7	100	9
cortisone acetate	2.37	95	85	10	78	17	66	29	89	6
cytarabine	−2.42	14	0	14	7	7	4	10	0	14
dacarbazine	−0.20	5	27	22	36	31	21	16	0	5
debrisoquine	0.14	85	45	40	27	58	42	43	57	28
deferiprone	−0.74	33	4	29	15	18	41	8	11	22
deflazacort	2.71	40	75	35	72	32	76	36	63	23
demeclocycline	−0.28	54	53	1	58	4	60	6	49	5
dicamba	2.41	99	79	20	80	19	79	20	56	43
didanosine	−1.26	3	25	22	10	7	25	22	51	48
diflunisal	3.63	99	90	9	94	5	83	16	77	22
dihydroergotamine	2.91	93	85	8	94	1	95	2	100	7
diltiazem	2.58	81	96	15	95	14	72	9	83	2
doxazosin	1.80	95	95	0	82	13	97	2	51	44
doxepin	2.79	80	75	5	88	8	71	9	39	41
encainide	3.76	70	91	21	89	19	82	12	94	24
etidocaine	2.82	94	89	5	63	31	67	27	79	15
etilefrin	−0.58	25	51	26	19	6	31	6	14	11
everolimus	3.32	74	90	16	66	8	100	26	78	4
fendiline	4.22	95	86	9	96	1	100	5	100	5
fenoprofen	3.05	99	95	4	98	1	94	5	100	1
fexofenadine	4.25	65	90	25	95	30	100	35	94	29
flunitrazepam	2.31	79	88	9	91	12	75	4	80	1
flurbiprofen	3.04	99	96	3	98	1	91	8	47	52
flutamide	2.91	90	73	17	84	6	82	8	86	4
fosfomycin	−1.11	2	20	18	7	5	29	27	24	22
fosinoprilat	2.25	98	84	14	79	19	76	22	64	34
galantamine	1.29	18	21	3	64	46	47	29	42	24
gemcitabine	−0.79	2	8	6	12	10	24	22	52	50
gentamicin	−3.64	20	24	4	8	12	0	20	0	20
gliclazide	1.02	91	86	5	81	10	71	20	64	27
glufosinate ammonium	−2.74	1	2	1	6	5	19	18	0	1
guanethidine	−0.38	5	3	2	15	10	1	4	0	5
hexobarbital	1.74	48	51	3	54	6	53	5	48	0
hydromorphone	1.59	7	21	14	24	17	48	41	44	37

Table 3 (Continued)

compound	CSLogP[a]	PBexp[b]	PBpred[c] (ANN)	AE[d]	PBpred (kNN)	AE	PBpred (MLR)	AE	PBpred (SVM)	AE
hydroxychloroquine	1.77	50	67	17	75	25	64	14	68	18
ibomal	1.54	34	51	17	47	13	46	12	14	20
idebenone	2.69	90	89	1	86	4	71	19	71	19
imipramine	4.70	88	95	7	95	7	86	2	92	4
indoprofen	2.72	99	96	3	97	2	93	6	79	20
isoniazid	−0.72	3	17	14	13	10	28	25	0	3
isosorbide mononitrate	−0.29	3	16	13	14	11	23	20	3	0
isradipine	3.63	97	90	7	92	5	85	12	55	42
itraconazole	3.93	100	90	10	75	25	100	0	100	0
lamivudine	−1.48	36	8	28	6	30	18	18	10	26
letrozole	3.31	60	73	13	63	3	59	1	42	18
levocabastine	3.01	55	89	34	90	35	85	30	79	24
levofloxacin	−0.40	27	25	2	34	7	52	25	37	10
levorphanol	3.15	40	63	23	90	50	64	24	47	7
liothyronine	1.36	98	92	6	82	16	71	27	73	25
lisinopril	−2.47	10	23	13	23	13	45	35	0	10
lomefloxacin	−0.47	12	31	19	24	12	47	35	22	10
lopinavir	5.35	99	87	12	97	2	100	1	100	1
lysergide	2.32	90	74	16	62	28	71	19	65	25
medroxyprogesterone	3.14	94	92	2	91	3	77	17	83	11
mefruside	1.56	65	79	14	51	14	69	4	32	33
meloxicam	2.03	100	97	3	75	25	84	16	71	29
melperone	3.57	50	60	10	75	25	67	17	78	28
meprobamate	0.76	20	23	3	38	18	21	1	0	20
meptazinol	2.83	27	37	10	58	31	53	26	50	23
methacycline	−0.15	84	60	24	61	23	60	24	51	33
methadone	3.21	84	88	4	85	1	72	12	100	16
methohexital	1.62	77	49	28	54	23	64	13	45	32
metildigoxin	1.77	15	37	22	38	23	69	54	44	29
metronidazole	−0.23	10	16	6	34	24	27	17	7	3
mexiletine	2.03	65	50	15	51	14	60	5	44	21
mibefradil	5.08	99	90	9	91	8	96	3	100	1
midazolam	2.85	96	91	5	82	14	78	18	89	7
montelukrast	5.28	99	98	1	83	16	100	1	100	1
mupirocin	2.88	96	91	5	88	8	71	25	75	21
nafarelin	5.74	80	50	30	31	49	86	6	23	57
naloxone	1.77	45	34	11	24	21	60	15	69	24
nefazodone	3.12	99	98	1	91	8	86	13	100	1
netilmicin	−3.75	10	23	13	7	3	0	10	0	10
nicardipine	4.59	98	98	0	94	4	91	7	64	34
nicorandil	−0.69	25	25	0	21	4	27	2	6	19
niridazole	0.56	85	38	47	17	68	39	46	39	46
nitrofurantoin	−0.21	63	49	14	21	42	33	30	32	31
nitroglycerin	1.18	60	55	5	55	5	43	17	27	33
nizatidine	0.08	29	3	26	18	11	40	11	0	29
norethisterone	2.75	80	90	10	86	6	75	5	90	10
norfloxacin	−0.15	15	39	24	42	27	52	37	0	15
octreotide	−0.51	65	37	28	74	9	57	8	27	38
olanzapine	3.14	93	84	9	90	3	87	6	93	0
ornidazole	0.13	15	22	7	36	21	35	20	20	5
oxitropium	3.95	7	7	0	20	13	25	18	26	19
paclitaxel	2.46	65	92	27	97	32	94	29	81	16
paramethadione	0.67	0	16	16	18	18	35	35	30	30
paroxetine	2.33	95	99	4	92	3	59	36	69	26
hepe	2.82	100	77	23	76	24	94	6	38	62
penicillin 13	2.07	66	70	4	65	1	58	8	49	17
penicillin 20	2.54	94	78	16	84	10	63	31	69	25
penicillin 24	3.12	94	92	2	94	0	88	6	100	6
penicillin 27	−0.05	60	51	9	35	25	48	12	94	34
penicillin 28	0.28	55	53	2	52	3	55	0	44	11
penicillin 30	−0.04	26	40	14	49	23	48	22	42	16
penicillin 31	−1.51	12	18	6	22	10	28	16	8	4
penicillin 38	2.50	62	82	20	79	17	76	14	61	1
penicillin 40	3.09	83	91	8	90	7	84	1	77	6
penicillin 45	2.64	65	87	22	80	15	74	9	63	2
penicillin 46	1.14	60	73	13	75	15	60	0	54	6
penicillin 48	2.21	82	79	3	77	5	72	10	49	33
penicillin 57	2.80	96	89	7	94	2	82	14	81	15
penicillin 60	1.50	86	68	18	78	8	62	24	59	27
penicillin 62	2.28	80	84	4	80	0	69	11	64	16
penicillin 70	3.02	97	88	9	90	7	88	9	48	49
penicillin 74	2.10	90	74	16	77	13	71	19	77	13
pentachlorophenol	5.08	99	98	1	80	19	100	1	92	7
perphenazine	3.65	92	99	7	92	0	91	1	45	47
phenformin	0.79	16	38	22	28	12	39	23	8	8
phentolamine	−1.22	54	52	2	57	3	41	13	71	17

Table 3 (Continued)

compound	CSLogP [a]	PBexp[b]	PBpred[c] (ANN)	AE[d]	PBpred (kNN)	AE	PBpred (MLR)	AE	PBpred (SVM)	AE
pirenzepine	0.60	10	43	33	66	56	39	29	52	42
piroxicam	1.28	99	91	8	68	31	79	20	84	15
pravastatin	4.41	48	97	49	93	45	97	49	64	16
praziquantel	1.19	83	78	5	88	5	62	21	65	18
prednisolone	1.46	88	72	16	75	13	61	27	72	16
prenylamine	4.29	97	86	11	96	1	100	3	100	3
probenecid	2.28	90	86	4	94	4	88	2	100	10
procainamide	1.10	17	11	6	49	32	37	20	19	2
propoxyphene	3.75	75	89	14	90	15	78	3	94	19
raloxifene	4.50	97	91	6	92	5	100	3	100	3
reproterol	−2.10	50	51	1	7	43	11	39	21	29
ritonavir	4.74	100	92	8	89	11	100	0	100	0
ropinirole	2.57	40	69	29	51	11	59	19	62	22
ropivacaine	2.39	94	88	6	48	46	59	35	63	31
sertraline	3.35	99	89	10	93	6	96	3	100	1
sodium cromoglicate	1.73	65	71	6	69	4	100	35	70	5
spironolactone	2.39	98	77	21	68	30	74	24	100	2
sulfamethoxazole	0.56	63	90	27	85	22	70	7	68	5
sulfamethoxypyridazine	0.77	75	90	15	84	9	89	14	87	12
sulindac	3.90	93	99	6	97	4	100	7	100	7
sulpiride	0.66	25	10	15	30	5	44	19	56	31
tacrine	2.01	55	80	25	76	21	68	13	70	15
tacrolimus	3.59	87	88	1	58	29	84	3	83	4
tebufelone	5.21	100	97	3	93	7	87	13	89	11
terazosin	2.09	91	89	2	75	16	78	13	86	5
terbutaline	−0.67	21	38	17	32	11	26	5	37	16
tetrazepam	3.50	70	94	24	79	9	89	19	50	20
tianeptine	1.63	95	87	8	81	14	94	1	100	5
ticarcillin	0.64	58	52	6	69	11	49	9	51	7
tilidine	3.05	25	82	57	67	42	70	45	60	35
tiludronic acid	−0.03	90	62	28	18	72	59	31	54	36
timolol	1.33	10	22	12	21	11	67	57	42	32
tocainide	0.83	13	33	20	34	21	54	41	32	19
tolamolol	1.30	91	61	30	48	43	49	42	47	44
tolbutamide	1.81	96	74	22	70	26	80	16	64	32
tolfenamic acid	4.81	100	98	2	99	1	100	0	92	8
topiramate	0.59	15	36	21	51	36	34	19	55	40
tramadol	2.20	13	30	17	43	30	44	31	0	13
troglitazone	2.53	100	95	5	93	7	75	25	100	0
tubocurarine	5.94	45	62	17	85	40	75	30	28	17
valacyclovir	−0.52	18	18	0	10	8	22	4	0	18
valsartan	2.95	95	96	1	90	5	100	5	94	1
vigabatrin	−2.85	0	1	1	6	6	17	17	0	0
vinorelbine	2.54	85	75	10	91	6	98	13	79	6
zanamivir	−3.31	5	10	5	15	10	0	5	0	5
zolpidem	2.80	93	87	6	77	16	86	7	72	21
zonisamide	1.08	40	84	44	57	17	62	22	16	24
zopiclone	1.45	45	31	14	69	24	45	0	21	24

[a] Predicted logP (CSPredict v2.0.3.1, ref 18). [b] Experimental % protein binding. [c] Predicted % protein binding by the indicated model. [d] Absolute error calculated as |predicted − experimental|

14%, 79%, and 9% of compounds in the range from −3 to 0, 0 to 5, and 5 to 7, respectively. Seventeen descriptors (in bold in Table 4) were found in two or more models. Seven descriptors in the ANN models were found in both the MLR and SVM models, five in common with the kNN, and three or more identical indices were common among the kNN or the MLR or SVM models. Two descriptors in the top 10 that are common to all models are logP and SallNp (sum of E-states for permanently charged nitrogens; see Tables 5−8). SCarom and e1C3N3 were next in commonality, found in three of the four models. As indicated in Tables 7 and 8, SCarom is the sum of E-state values of all aromatic carbon types; e1C3N3 is the sum of the bond E-states for a sp^3 carbon (>CH-) single-bonded to a tertiary nitrogen atom (>N-) where the letter, e, signifies a bond-type.

Of some additional interest is the frequency of the descriptors. As seen by the frequency and ranking values in Table 4, the important ranked topological descriptors are independent of their frequency percentages. The rank and frequency values are uncorrelated: $r^2 = 0.03$, averaged over the four models. Furthermore, no model has a majority of global descriptors (a frequency >90%). ANN, kNN, MLR, and SVM had two, eight, four, and four global descriptors, respectively. The average frequency for nonglobal descriptors is 36%, about one in every three compounds; however, 40% of these nonglobal indices have an average occupancy of 14.6% or about one in every seven compounds. This information indicates, in the optimization process, models are selecting descriptors which represent very specific structural attributes.

Two models, ANN and MLR, were amenable to a mean trend analysis. A descriptor trend points out the relationship between the change in descriptor value and the resulting change in calculated protein binding value. In the procedure for determining a descriptor trend, the descriptor value is incremented 100% (up/down 50%) of its range in 10 evenly spaced intervals, while all other indices are held constant. The resulting set of computed property values is plotted against the changed descriptor values. These plots were examined for trends. In general, some plots exhibit increasing or decreasing relationships which are nearly but not exactly linear. Others indicate a nonlinear relation with

Table 4. Ranking of Top 20 Descriptors in QSAR Models[a]

ANN	FQ[b]	Trend[c]	kNN	FQ[b]	MLR	FQ[b]	Trend[c]	SVM	FQ[b]
CSLogP	808	+	**CSLogP**	808	**CSLogP**	808	+	**CSLogP**	808
Ssp3N	177	-	**SallNp**	18	**SallNp**	18	-	**AromMol**	646
SallNp	18	-	e1N2S4dd	38	**Gmin**	808	-	SaaCH	628
SHBint2	471	+	dxp4	801	**xch6**	725	+	**SCarom**	656
SHCarOH1	225	+	**dxv0**	807	Ssp2OH	232	+	**eaC2C2a**	558
SCarOH1	225	-	dxvp8	785	**Ssp3N**	177	-	SsssCH	425
SHBint4	195	-	**SCarom**	646	SaasC	642	-	SddssS	54
eaC2C3s	600	+	xp4	801	**eaC2N2a**	95	-	e1C3S4da	56
Gmin	808	-	xp10	728	xvch6	725	-	SHArom	628
SotArom	223	+	dxv2	807	SdsCH	162	+	**SallNp**	18
SPheOH1	74	-	dxvp7	791	**SaaN**	166	+	SsFCl	186
e1C3N3	223	-	SaasN	55	e1C3N1d	232	-	**eaC2C3s**	600
SCarom	646	+	e1C3O1d	41	SsCl	139	+	dx1	808
Ssp3NH	81	-	NArom	197	**SOAmide2**	37	+	**xch6**	725
e2C3O1s	565	-	**SaaN**	166	xvp8	782	+	SssssC	310
SsssN	425	variable	xvch9	151	dx2	807	-	**SOAmide2**	37
xch10	195	+	dxvp6	797	**SHBint4**	195	-	dxv1	808
eaC2C2a	558	-	SHvin	162	SArNH2	54	-	**dxv0**	807
eaC2N2a	95	-	**e1C3N3**	223	**xch10**	195	+	**e1C3N3**	223
SsNH2	174	-	**Ssp3NH**	81	**AromMol**	646	+	**SsNH2**	174

[a] Grayed-out descriptors signify descriptors found in two or more models; Boxed-out descriptor indicate a homologous descriptor found in two or more models. [b] Frequency (FQ) is the number of compounds in the 808 compound train/test sets with a nonzero value for the given descriptor. [c] A positive trend indicates % protein binding will generally increase with an increase in the descriptor value. A negative trend indicates the inverse effect. Trends marked as variable may have either a positive or negative effect on predicted %PB depending on the value of other descriptors. The ANN and MLR mean trends were evaluated by holding all other descriptor values constant and varying the given descriptor value by 50% up and down to determine the directional change in %PB.

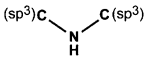
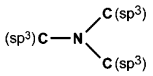
Table 5. Definitions of Important Global Descriptors and Binary Indicators Including Connectivity Indices[a]

index	name	information encoded
CSlogP	predicted LogP (CSPredict v2.0.3.1 [18])	lipophilicity
TPSA	total polar surface area	topological surface area of hydrogen bonding atoms (Ertl method)
Gmin	minimum atom level E-State value	atom associated with a site of electrophilic attack
xp4	chi simple path 4	skeletal complexity and details of
xp10	chi simple path 10	arrangement of branching
xch6	chi simple chain 6	complexity of large and fused ring systems
xch10	chi simple chain 10	and degree of ring substitution
xv1	chi valence 1	molecular weight and volume
	connectivity index for path of length 1	presence of heteroatoms
xvp8	chi valence path 8	skeletal complexity, branching, and heteroatoms
xvpc4	chi valence path-cluster 4	adjacency of branching
xvch6	chi valence chain 6	complexity of large and fused ring systems
xvch9	chi valence chain 9	and degree of ring substitution and
xvch10	chi valence chain 10	presence of heteroatoms
dx1	chi simple difference 1	branching independent of molecular size
dx2	chi simple difference 2	
dxp4	chi simple difference path 4	branching independent of molecular size
dxp5	chi simple difference path 5	complexity of branching
dxv0	chi valence difference 0	branching independent of molecular size
dxv1	chi valence difference 1	presence of heteroatoms
dxv2	chi valence difference 2	
dxvp6	chi valence difference path 6	branching independent of molecular size
dxvp7	chi valence difference path 7	complexity of branching
dxvp8	chi valence difference path 8	presence of heteroatoms
AromMol	indicator variable for the presence of an aromatic group	
SOAmide2	indicator variable for the presence of a secondary sulfonamide	
NArom	indicator variable for the presence of an aromatic nitrogen group	

[a] Some specific information encoded by individual connectivity indices is listed after the forst three entries. In models of large diverse datasets, however, the principle impact of the molecular connectivity indices is not expressed in the form of a specific independent contribution for each index but rather as collectively encoded information useful in differentiating among the different types of skeletal scaffolds that are present in the training set.

a maximum or minimum; still others show a sigmoidal-like relationship. An interesting result was found among the trends associated with seven descriptors (logP, SallNp, Gmin, Ssp3N, SHBint4, xch10, and eaC2N2a) common to both ANN and MLR

Table 6. Definitions of Important Atom-type and Hydrogen Atom-type Descriptors in the Four Models

Index	Description	Illustration
SsssCH	Sum of atom level E-state values of all methine carbon atoms in the molecule	
SsssC	Sum of atom level E-state values of all quaternary carbon atoms in the molecule	
SdsCH	Sum of atom level E-state values of all vinyl carbon atoms in the molecule	
SaaCH	Sum of atom level E-state values of unsubstituted aromatic carbon atoms in the molecule	
SaasC	Sum of atom level E-state values of all substituted aromatic carbon atoms in the molecule	
SHvin	Sum of atom level hydrogen E-state values of all hydrogen atoms in the molecule on vinyl carbons	
SsNH2	Sum of atom level E-state values of all primary amine nitrogen atoms in the molecule	
SsssN	Sum of atom level E-state values of all tertiary amine nitrogen atoms in the molecule	
Ssp3NH	Sum of E-states values of all secondary amine in the molecule that are bonded to two sp3 carbons	
Ssp3N	Sum of E-states values of all tertiary amine in the molecule that are bonded to three sp3 carbons	
SaaN	Sum of atom level E-states values of all pyridine nitrogen atoms in the molecule	
SaasN	Sum of atom level E-states values of all substituted pyrole nitrogen atoms in the molecule	
SArNH2	Sum of atom level E-states values of all primary aniline nitrogen atoms in the molecule	
Ssp2OH	Sum of atom level E-states values of all –OH oxygen atoms in the molecule that are bonded to a sp2 carbon atom	
SddssS	Sum of atom level E-states values of all ddss sulfur atoms	
SsCl	Sum of atom level E-state values for all chlorine atoms in the molecule	

models. All seven indices exhibited identical positive (or negative) trends in these two models: This finding may underscore the importance of these descriptors. It is apparent from the trends that hydrophilicity, aromaticity, presence of a ring structure, and the presence and bonding state of amines play important stereochemical roles in serum plasma protein binding of the compounds.

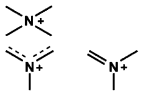
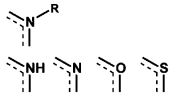
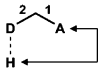
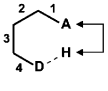
Discussion

To provide a basis for relating many of the most-important descriptors found in the four QSAR models to physicochemical properties of the compounds in this study, Table 9 presents a profile of ionization states. The percentage of compounds is given for each of four groups: acid, base, neutral, zwitterionic, and permanently charged compounds (e.g., quaternary nitrogen atoms). Ionization state applies to pH 7.4 for the train/test and validation datasets. Approximately 90% of the 808 members in the train/test set are evenly represented by acids, bases, and neutral entities. Approximately 7% of the remaining members

Table 7. Definitions of Important Bond-type Descriptors in the Four Models

Index	Description	Illustration
eaC2C2a	Sum of the bond E-state values for unsubstitued aromatic carbons	
eaC2C3s	Sum of bond E-state values for aromatic carbons with one substituent group	
e1C3N3	Sum of bond E-State values between tertiary amines and >CH– carbon atoms	
e1C3N1d	Sum of bond E-State values between imine nitrogen atoms and >CH– carbon atoms	
eaC2N2a	Sum of the bond E-state values for aromatic bond between carbon and nitrogen	
e2C3O1s	Sum of the bond E-state values for double bonds between carbonyl oxygen and >C– carbon atoms	
e1C3O1d	Sum of the bond E-state values for single bond between alcohol oxygen and =C– carbon atoms	
e1C3S4da	Sum of the bond E-State values for single bond between aromatic carbon and ddss sulfur atom	
e1N2S4dd	Sum of the bond E-State values for single bond between secondary amine nitrogen atoms and and ddss sulfur atom	

Table 8. Definitions of Important Group-type and Single-Atom Descriptors in the Four Models

Index	Description	Illustration
SCarom	Sum of E-state values for all substituted and unsubsituted aromatic carbon atoms	
SsFCl	Sum of atom E-state values for all fluorine and chlorine atoms in the molecule	
SHArom	Sum of HE-state values for all hydrogen atoms attached to aromatic carbon atoms	
SallNp	Sum of E-state values for all quaternary, pyridinium and aminium nitrogen atoms in the molecule	
SotArom	Sum of atom E-state values for all heteroaromatic atoms in the molecule (pyridine, pyrole, substituted pyrole, thiophene or furan)	
SPheOH1	Largest E-state value of all phenolic oxygen atoms In the molecule	
SHCarOH1	Largest HE-state value in the molecule of all hydrogen atoms bonded to carboxyl oxygens (acidic oxygen)	
SHBint2	Largest product of E-state and HE-state from all acceptor(A) and donor(D) pairs in the molecule separated by 2 skeletal bonds	
SHBint4	Largest product of E-state and HE-state from all acceptor(A) and donor(D) pairs in the molecule separated by 4 skeletal bonds	

are zwitterions and 3% with a permanent positive charge. The ANN descriptors clearly indicate the importance of ionizable

Table 9. Characterization of Ionizable Compounds in Train/Test and Validation Sets at pH 7.4

compounds	number	percent[b]	MW[c]	CSLogP[d]
		Training Set		
acid[a]	213	26.4%	369.7	1.65
base	249	30.8%	347.0	2.22
neutral	268	33.2%	350.3	1.66
zwitterionic	56	6.9%	406.8	−0.01
permanent (+)[e]	22	2.7%	460.5	
		Validation Set		
acid[a]	46	23.0%	371.4	2.08
base	63	31.5%	376.2	2.21
neutral	67	33.5%	354.8	1.55
zwitterionic	21	10.5%	343.1	−0.38
permanent (+)[e]	3	1.5%	491.9	

[a] Acid, base, zwitterion, or neutral compound determined by estimated pK_a values for the compound's ionizable groups using the CSpK_a predictor from ChemSilico LLC. [b] Percent of train or test set in each ionization group. [c] Average molecular weight for ionization class. [d] Average predicted logP for ionization class. [e] Compounds with permanent positive charge are quarternary amines, pyridinium, or animinium compounds.

and formally charged groups by including two descriptors for carboxylic acids [SHCarOH1, SCarOH1], five for bases [Ssp3N and Ssp3NH (N bonded only to sp^3 carbons), SsNH2, and SsssN (bonded to any type carbon), e1C3N3], and finally SallNp (a permanent positive charge on nitrogen). The negative trend found for SallNp in both ANN and MLR is consistent with the general tendency for compounds with a formal positive charge to have low protein binding. Tables 5−8 give a listing and description of many of the most important descriptors in these four QSAR models.

The kNN, MLR, and SVM models had no important descriptors that represent carboxylic acid. However, a binary indicator (SOAmide2) for the presence/absence of sulfonyl secondary amide was included in the MLR and SVM models where the amine may likely be partially or fully deprotonated at pH 7.4, hence, acidic. For the SVM model, two indices for carboxylic acids occur but they are not in the top 20 descriptors. The ANN model had four amine descriptors (Ssp3N, e1C3N3, Ssp3NH, SsNH2) in the top 20 ranked members reflecting the importance, in a negative sense, of amines in protein binding for this model. In contrast, there were only two or less amine descriptors among the important members for kNN, MLR, or SVM models; while, on the other hand, all of these three models did include SallNp. Aromaticity is another important physicochemical property of these drugs; 81% of the 808 train/test dataset contain one or more aromatic rings (Table 1). In all four models, four or more aromatic descriptors were found as being important either as atom or bond-type E-states; i.e., SCarom (sum E-states for all aromatic carbon types) was included in the ANN, KNN, and SVM models. The bond-type E-States, eaC2N2a (E-State for bond between aromatic C and N), eaC2C3s (bond E-State between two aromatic carbons, one with a substituent group), and SaaN (sum of E-States for aromatic nitrogen atoms) were included in one or more of the models. Therefore, in varying degrees the models reflect the importance of both ionizable groups and aromaticity. In some cases, analogous descriptors are included in different models (see Table 4). For example, eaC2C3s found in the ANN and SaasC in the MLR model are analogous for both aromatic atoms and bonds with substituents.

Let us now consider how these important topological descriptors, the 19 together with logP in Table 4, relate to QSAR variables found in previous studies on serum plasma proteins or HSA binding. Recent efforts over several years to develop QSAR models for serum protein or human serum albumin binding of drug or drug-like compounds have focused on moderately sized datasets[13,16,27−33] with the exception of one large chemometric analysis.[33] In all of these studies, calculated logP was found to be the most important descriptor in QSAR models involving other physicochemical parameters. By contrast, LogD was found to have little or no correlation to HSA binding.[9,31]

In one particular study by Colmenarejo[29] using HSA-HPLC affinity column chromatography data (based on immobilized albumin) on 94 drugs, xch6 (a ring molecular connectivity descriptor) was found in two models, ranked second behind logP in importance, followed by three electronic and surface area descriptors. The purpose of this model was to predict the retention index, logk(HSA), which is considered to be directly correlated to %PB. Using the reported HPLC data from this latter study, Hall et al.[32] developed a QSAR-MLR model constructed only on topological indices without the use of logP. With only five topological indices employed, the 84 drugs in the training set gave an $r^2 = 0.77$ for all 84 compounds. In the Colmenarejo model, only 79 compounds were included because five were found to be large outliers, for the diminished data set $r^2 = 0.78$. With a 10 compound validation set and the five topological indices, the Hall QSAR model gave an MAE = 0.31 with no large residuals. Of the five descriptors found (SCarom, xch6, SsFCl, SHCsats, and SsOH), the first three are found among the most important variables in our QSAR protein binding models; the last two of secondary importance.

The structure descriptors found to be important in the ANN model may be compared to fragments from a chemometric model involving affinity binding to the 3A subdomain of HSA.[33] The model was based on contributions of 74 fragments for a data set of 889 compounds. These data was composed of 232 compounds with $K_d < 10^{-3}$ M, considered as active binders, and 657 compounds considered inactive (poor binders with $K_d > 10^{-3}$ M). Table 10 summarizes the comparison of relevant ANN descriptor trends with fragments counts. Eighteen individual topological descriptors out of 33 from the ANN model could be mapped specifically to drug fragments found important for binding to the 3A subdomain. More notably all 18 topological descriptors had their predominant trends in the same direction as trends based on the sign of the weights assigned to fragments in the 3A domain binding study model. Minor exceptions are also evident in Table 10 with respect to trends based on the descriptors (signified by + and − signs) versus trends assigned to fragments. Nonetheless, it is gratifying that such an agreement exists among trends in the ANN model and those coming from the fragment approach, considering the following:

(a) The fragment model is based only on 3A subdomain affinity binding whereas the ANN train set involved intact HSA.

(b) The fragment model is based on a linear regression model rather than the nonlinear machine-learning algorithm with its totally different selection processes for the independent variables.

(c) The fragment model uses affinity dissociation constants, K_d, to develop the fragment model rather than %PB values as used here in the QSAR ANN model.

It is also worth noting in the MLR %PB model, 10 out of 14 of the highly ranked topological indices have the same trend as the corresponding fragments. Although a trend analysis is not available for the kNN and SVM models, 15 out of 28 and 23 out of 60 topological indices, respectively, have bond-type or atom-type E-states corresponding to fragments found in the list of 74 fragments.[33]

Unfortunately, no major study specific to Site 1, the warfarin site, in the IIA subdomain of HSA has been conducted as with

Table 10. Comparison of Trends for ANN Descriptors and Fragment Based QSAR Model

ANN descriptor	trend[a]	fragment-type[b]	trend[c]
Ssp3N	−	amines	−
Ssp3NH	−	(−NH_2, >NH, >N−)	
SsNH2	−	piperidine	+
e1C3N3	−/+	pyrrolidine	+
e1C2N2	−/+	morpholine	+
e2C3O1s	−	amides	−
SHBint2	±	ketone	−
		aldehyde	−
		ester	−
		ureas	−
		hydantoin	−
		pyrimidinone	−
		pyrrolinone	−
		piperdinone	−
		carbamates	−
SallNp	−	quaternary amine	−
SPheOH1	−/+	phenol	−
SCarOH1	+	−CO_2H	+
SHBint2	±		
SCarom	+	non-N-heteroaromatics	+
SotArom	+	naphthalene	
xc*h*10	+	phenyl	
eaC2C3s	±		
SCarom	+	N-heteroaromatics	±
SotArom	+		
eaC2C3s	±		
eaC2N2a	+		
SssS	−	thioether	−
SssO	−	ether	−
SssCH2	+	methylene	+

[a] Trend sign is the mean and ± or −/+ to indicate the top sign is dominate and lower < 50% of train set of compounds. [b] Data from ref 33. [c] Fragment + trend signifies presence of fragment increased binding to the 3A-domain of HSA or the reverse and ± indicates exceptions.

Site 2 binders. Nonetheless, 14 compounds contained in our 808 train set are listed as binders[9] to Site 1 and provide some limited insight into trends of the ANN model's top 20 indices in regard to Site 1 binders. All of these 14 compounds contain at least one aromatic ring, five with heteroaromatic rings, and four with two-membered, fused ring systems. The 14 compounds were ceftriaxone, chlorpropamide, dicoumarol, etodolac, furosemide, indomethacin, oxyphenbutazone, phenylbutazone, phenytoin, sulfathiazole, suprofen, thyroxine, urapidil, and warfarin. Looking at compounds where at least 6 out 14 compounds contained the same descriptor, left 10 suitable topological indices to examine from 20 important ANN descriptors. SsNH2, e2C3O1s, Gmin, and SCarOH1 all showed a negative trend indicating presence of a carboxylic acid or amine or electron withdrawing group signified by Gmin. Each one resulted in diminished binding at Site 1. Not too surprising was the positive trends for bond and E-States related to aromaticity; i.e., eaC2C2a, eaC2C3s (branched aromatic carbon), SCarom, and SotArom. SHBint4 and SHBint2 both showed a positive trend. The latter descriptor suggests that amides, contained in seven compounds, may aid in binding as well as H-acceptor/donor groups separated by four bonds as in the case of SHBint4. The most surprising finding among these 14 Site 1 binders was compound lipophilicity. It appeared to have a neutral effect. The computed logP ranged from 0.38 to 3.98 for the 14 compounds but the trend was flat with large changes in logP for these compounds. A similar trend analysis with the MLR model also showed this same neutral influence of logP on %PB values of these Site 1 binders. This is in stark contrast to the pronounced, positive dependency on logP for Site 2 binders. Since the number of Site 1 compounds was limited here, no firm conclusion can be draw on this observed lack of logP dependency on %PB values until a much larger number of Site 1 binders are identified. Of two remaining identical descriptors, Gmin and SHBint4, found in both the MLR and ANN models, their trends in MLR model were same as the trends observed in the ANN trend analysis. These limited results on the trends for site I binders for various descriptors in the ANN are mainly consistent with what is known about Site 1 binders. They are bulky heterocyclic compounds.[7]

Conclusions

Results in this study clearly demonstrate the usefulness of topological descriptors in combination with logP to tackle the difficult problem of human serum protein binding prediction. The statistical results on the external validation set indicate that the ANN model is useful for prediction of new chemical entities. These current results are consistent with others obtained earlier using the structure−information approach for aqueous solubility, human intestinal absorption, and Ames mutagenicity.[34−36]

Of the four models presented in this study, ANN and kNN are the two most robust ensemble models with a distinct difference; the kNN model employed about 50% more connectivity indices and 50% fewer atom-type E-state descriptors than the ANN model. Such differences may reflect, in part, establishment of nonlinear relationships between protein binding and descriptors by the ANN learning algorithm as compared to molecular similarity matching by the kNN approach. Nonetheless, an analysis of the structure descriptors found in the ANN and kNN models provides a basis for the chemist to develop structure modifications during the drug design process. The role of nitrogen-containing compounds, acids, aromatic entities (atom-type and bond-type E-State descriptors), and skeletal ramification (molecular connectivity chi indices) are all included in the structure−information. For each descriptor the trend with respect to protein binding is reported and can be used as an indication of the impact of structure modification on predicted protein binding.

The structure descriptors included in the ANN model were compared to a list of fragments found important for ligand binding to the IIIA subdomain of albumin. Analysis of the IIIA fragment trends with those of important modeling descriptors in this study clearly demonstrates the structure−information content of topological indices can be related directly back to fragments. This information complements what we know about fragment contributions as they relate to topological indices and makes the topological structure descriptors that much more relevant to the chemist. Use of fragments suffers from the problem that an individual fragment may or may not correspond to an active modulator depending on compound. Further, compounds with missing fragments cannot be predicted by such a method. Topological structure descriptors generally do not suffer from these problems.[36]

Supporting Information Available: A listing by name, chemical formula, molecular weight, and experimental percent binding to serum proteins is provided. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Tawara, S.; Matsumoto, S.; Matsumoto, Y.; Kamimura, T.; Goto, S. Structure-binding Relationship and Binding Sites of Cephalosporins in Human Serum Albumin. *J. Antibiot.* **1992**, *45*, 1346−1357.
(2) Sharples, D. Competition for Plasma Protein Binding Sites Between Phenothiazine Tranquillizers and Iminodibenzyl Antidepressants. *J. Pharm. Pharmacol.* **1975**, *27*, 379−381.
(3) Schon, A.; del Mar Ingaramo, M.; Freire, E. The Binding of HIV-1 Protease Inhibitors to Human Serum Proteins. *Biophys. Chem.* **2003**, *105*, 221−230.

(4) Kuchinskiene, Z.; Carlson, L. A. Composition, Concentration, and Size of Low-Density Lipoproteins, and Subfractions of Very Low-Density Lipoproteins From Serum of Normal Men and Women. *J. Lipid Res*. **1982**, *23*, 762−769.
(5) He, X. M.; Carter, D. C. Atomic Structure and Chemistry of Human Serum Albumin. *Nature* **1992**, *358*, 209−215.
(6) Curry, S.; Mandelkow, H.; Brick, P.; Franks, N. P. Crystal Structure of Human Serum Albumin Complexed with Fatty Acid Reveals an Asymmetric Distribution of Binding Sites. *Nat. Struct. Biol*. **1998**, *5*, 827−835.
(7) Petitpas, I.; Ananyo, A.; Bhattacharya, A. A.; Twine, S.; East, M.; Stephen, C. Crystal Structure Analysis of Warfarin Binding to Human Serum Albumin. *J. Biol. Chem*. **2001**, *276*, 22804−22809.
(8) Sugio, S.; Kashima, A.; Mochizuki, S.; Noda, M.; Kobayashi, K. Crystal Structure of Human Serum Albumin at 2.5 Å Resolution. *Protein Eng*. **1999**, *12*, 439−446.
(9) Kratochwil, N. A.; Huber, W.; Mueller, F.; Kansy, M.; Gerber, P. R: Predicting Plasma Protein Binding of Drugs: A new approach. *Biochem. Pharmacol.* **2002**, *64*, 1355−1374.
(10) Paal, K.; Mueller, J.; Hegedus, L. High Affinity Binding of Paclitaxel to Human Serum Albumin. *Eur. J. Biochem*. **2001**, *268*, 2187−2191.
(11) Bhattacharya, A. A.; Curry, S.; Franks, N. P. Binding of the General Anesthetics Propofol and Halothane to Human Serum Albumin. *J. Biol. Chem*. **2000**, *275*, 38731−38738.
(12) Thummel, K. E.; Shen, D. D. In *Goodman and Gilman's The Pharmacological Basis of Therapeutics*, 10th ed.; Hardman, J. G., Limbird, L. E., Goodman, A. G., Eds.; McGraw-Hill: New York, 2001; pp 1924−2023.
(13) Dollery, C., Ed. *Therapeutic Drugs*, 2nd ed.; Churchill Livingstone: Edinburgh, 1999.
(14) Lacy, C. F.; Armstrong, L. L.; Goldman, M. P.; Lance, L. L. *Drug Information Handbook;* Lexi-comp, 2000.
(15) Moffat, A. C., Osselton, M. D., Widdop, B., Eds. *Clarke's Analysis of Drugs and Poisons*; Pharmaceutical Press: London-Chicago, 2004.
(16) Yamazaki, K.; Kanaoka, M. Computational Prediction of the Plasma Protein-Binding Percent of Diverse Pharmaceutical Compounds. *J. Pharm. Sci*. **2004**, *93*, 1480−1494.
(17) See the web site: www.rxlist.com.
(18) Topological structure descriptors are calculated as part of all ChemSilico products, based on the Molconn-Z software. CSLogP computes the octanol/water partition coefficient, ChemSilico LLC, Tewksbury, MA.
(19) Hall, L. H.; Hall, L. M. QSAR Modeling Based on Structure-Information for Properties of Interest in Human Health. *SAR QSAR Environ. Res.* **2005**, *16*, 13−41.
(20) MDL QSAR, v2., MDL Information Systems, San Leandro, CA.
(21) Devillers, J. Strengths and Weaknesses of the back-propagation neural network in QSAR and QSPR studies. In *Genetic Algorithms in Molecular Modeling;* Devillers, J., Ed., Academic Press: San Diego, CA, 1996; pp 1−24.
(22) Miller, A. *Subset Selection in Regression*, 2nd ed.; Chapman & Hall/CRC Press: Boca Raton, FL, 2002.
(23) Zheng, W.; Tropsha, A. Novel variable selection quantitative structure−property relationship approach based on the k-nearest-neighbor principle. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185−194.
(24) Sharaf, M. A.; Illman, D. L.; Kowalski, B. *Chemometrics*; John Wiley and Sons: New York, 1986.
(25) Shen, M.; Beguin, C.; Golbraikh, A.; Stables, L.; Kohn, H.; Tropsha, A. Application of Predictive QSAR Models to Database Mining: Identification and Experimental Validation of Novel Anticonvulsant Compounds. *J. Med. Chem*. **2004**, *47*, 2356−2364.
(26) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer-Verlag: New York, New York, 1995.
(27) Saiakhov, R.; Stefan, L. R.; Klopman, G. Multiple computer-automated structure evaluation model of the plasma protein binding affinity of diverse drugs. *Perspect. Drug Disc. Des.* **2000**, *19*, 133−155.
(28) Beaudry F.; Coutu, M.; Brown, N. K. Determination of drug-plasma protein binding using human serum albumin chromatographic column and multiple linear regression model. *Biomed. Chromatogr*. **1999**, *13*, 401−406.
(29) Colmenarejo, G.; Alvarez-Pedraglio, A.; Lavandera, J. L. Chemoinformatic models to predict binding affinities to human serum albumin. *J. Med. Chem.* **2001**, *44*, 4370−4378.
(30) Colmenarejo, G. *In silico* prediction of drug-binding strengths to human serum albumin. *Med. Res. Rev.* **2003**, *23*, 275−301.
(31) Valko, K.; Nunhuck, S.; Bevan, C. Abraham, M. H.; Reynolds, D. P. Fast gradient HPLC method to determine compounds binding to human serum albumin. Relationships with octanol/water and immobilized artificial membrane lipophilicity. *J. Pharm. Sci.* **2003**, *92*, 2236−2248.
(32) Hall L. M.; Hall, L. H.; Kier, L. B. Modeling drug albumin binding affinity with E-State topological structure representation. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 2120−2128.
(33) Hajduk P. J.; Mendoza, R.; Petros, A. M.; Huth, J. R.; Bures, M.; Fesik, S. W.; Martin, Y. C. Ligand binding to domain-3 of human serum albumin: A chemometric analysis. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 93−102.
(34) Votano, J. R.; Parham, M. E.; Hall, L. H.; Kier, L. B.; Hall, L. M. Prediction of Aqueous Solubility Based on Large Datasets using Several QSPR Models Utilizing Topological Structure Representation. *Chem., Biodiversity* **2004**, *1*, 1829−1841.
(35) Votano, J. R.; Parham, M. E.; Hall, L. H.; Kier, L. B. New Predictors for Several ADME/Tox Properties: Aqueous Solubility, Human Oral Absorption, and Ames Genotoxicity Using Topological Descriptors. *J. Mol. Diversity* **2004**, *8*, 385−397.
(36) Votano, J. R.; Parham M.; Hall, L. H.; Kier, L. B.; Oloff S.; Tropsha, A.; Xie, Q.; Tong, W. Three New Consensus QSAR Models for the Prediction of Ames Genotoxicity. *Mutagenesis* **2004**, *19*, 365−378.

JM051245V